A SYSTEM FOR DETECTION OF TRANSITION AND SPECIAL EFFECTS IN VIDEO

FIELD OF THE INVENTION

[0001] The invention relates to the field of multimedia technologies. More specifically, the invention relates to the detection of transition and special effects in videos.

DESCRIPTION OF THE RELATED ART

[0002] The act of detecting transition and special effects in video enables segmentation of video into its basic component, the shots. Typically a shot is considered an uninterrupted or "transition" free video sequence, such as a continuous camera recording. Video editing techniques may use any one of a number of effects to transition from one shot to another. These transition edit types include hard cuts, fades, wipes, dissolves, irises, funnels, mosaics, rolls, doors, pushes, peels, rotates, and special effects. Hard cuts are typically the most common transition effect in videos.

[0003] Automatic shot boundary detection techniques attempt to indicate where a transition effect occurs within an edited video stream. The complexity of detecting a shot boundary varies with the type of transition edit used. For example, hard cut, fade and wipe type edits generally require less complex detection techniques compared to dissolves type edits. This is because, in the case of hard cuts and fades, the two sequences involved are temporarily well-separated. Therefore, the detection technique used for hard cuts and fades are often determined by detecting that the video signal is abruptly governed by a new statistical process or that the video signal has been scaled by some mathematically well-defined and simple function (e.g. fade in, fade out).

[0004] Even in the case of wipes, the two video sequence involved in the transitions are well-separated at any time. This is typically not the case for a dissolve.

[0005] A dissolve is commonly defined as the superposition of a fading out and a fading in sequence. At any time, in regard to dissolves, two video sequences are temporally, as well as spatially intermingled. In order to employ a dissolve's definition directly for detection, the two sequences must be separated. Therefore there is a problem of two source separation.

[0006] For example, a dissolve sequence D(x, t) is defined as the mixture of two video sequences $S_1(x, t)$ and $S_2(x, t)$, where the first sequence is fading out while the second is fading in:

$$D(x,t) = f_1 \cdot S_1(x,t) + f_2 \cdot S_2(x,t) \text{ with } t \in [0,T]$$

[0007] Dissolve types are commonly cross-dissolves with

$$f_1 = \frac{T-t}{T}, t \in [0,T]$$
$$f_2 = \frac{t}{T}, t \in [0,T]$$

and additive dissolves with

$$f_1 = \begin{cases} 1 & \text{if } (t \le c_1) \\ \frac{T-t}{T-c_1} & \text{else} \end{cases}, t \in [0,T], c_1 \in]0,T[$$

$$f_2 = \begin{cases} \frac{t}{c_2} & \text{if } (t \le c_2) \\ 1 & \text{else} \end{cases}, t \in [0,T], c_2 \in]0,T[$$

[0008] In general, three different types of dissolves can be distinguished based on the visual difference between the two shots involved. Regarding a type one dissolve, the two shots involved have different color distributions. Thus, they are different enough such that a hard cut would be detected between them if the dissolve sequence were removed.

[0009] Regarding a type two dissolve, the two shots involved have similar color distributes which a color histogram-based hard cut detection algorithm would not detect. However, the structure between the images is different enough in order to be detectable by an edge-based algorithm. For example a transition from one cloud scene to another.

[0010] Regarding a type three dissolve, the two shots involved have similar color distributions and similar spatial layout. This type of dissolve is a special type of morphing.

[0011] Rule-based systems may be beneficial to achieve a computer vision and image understanding but only for simple problems. Existing shot detection methods can be classified as rule-based approaches. A main advantage of rule-based systems are that they usually do not require a large training set. Therefore, automatic shot boundary detection is normally attacked by a rule-based detection system, and not cast as a complex detection problem.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The accompanying drawings illustrate embodiments of the invention. In the drawings:

[0013] Figure 1 is a block diagram illustrating an overview of the training components according to one embodiment.

[0014] Figure 2 visualizes the various parameters of the transition generation synthesizer according to one embodiment.

[0015] Figure 3 illustrates a system overview of a transition detection system using a multi-resolution approach according to one embodiment.

[0016] Figure 4 illustrates a typical time series of the edge strength feature according to one embodiment.

[0017] Figure 5 illustrates the performance of the various features for pre-filtering according to one embodiment.

[0018] Figure 6 is a block diagram further illustrating the creation of the training set of block 200 according to one embodiment.

[0019] Figure 7 is a block diagram further illustrating the creation of the training and validation set of block 100 according to one embodiment.

DETAILED DESCRIPTION OF THE DRAWINGS

[0020] The present invention provides for detection of transition and special effects in videos. In the following description, numerous specific details are set forth to provide a thorough understanding of the invention. However, it is understood that the invention may be practiced without these specific details. In other instances, well-known protocols, structures and techniques have not been shown in detail in order not to obscure the invention.

[0021] The techniques shown in the figures can be implemented using code and data stored and executed on computers. Such computers store and communicate (internally and with other computers over a network) code and data using machine-readable media, such as magnetic disks; optical disks; random access memory; read only memory; flash memory devices; ASIC, DSP, electrical, optical, acoustical or other form of propagated signals (e.g., carrier waves, infrared signals, digital signals, etc.); etc. Of course, one or more parts of the invention may be implemented using any combination of software, firmware, and/or hardware.

[0022] One embodiment includes two components: a training system and a transition detection system. The training system includes a transition synthesizer. The transition synthesizer can create from a proper video database an infinite number of transition/special effect examples. In the remainder of the patent application we will use the dissolve transition as an the main example of a transition effect. It should be understood that this is not a restriction. The transition synthesizer is used to create a training and validation set of dissolves with a fixed scale (length) and a fixed location (position) of the dissolve center. These sets are then used to iteratively train an heuristically optimal classifier. For example, in one embodiment, the classifier is accomplished by pattern recognition and machine learning techniques.

[0023] Figure 1 is a block diagram illustrating an overview of the training components according to one embodiment of the invention. In block 100, the system creates a large set of synthetic training and validation patterns for selected transitions effects, then control passes to block 200. In block 200, the system performs iterative training of transition/effect detector and then control passing to block 300. In block 300, a fixed-scale and fixed-location transition detector is generated.

[0024] The significance that synthetic transitions may not be representative for real transitions, is minimal, because all transitions in real videos have been originally generated in exactly the same way. In one embodiment, the video database typically would consist of a diverse set of videos such as home videos, feature films, newscast, soap operas, etc. It serves as the source of video sequences for the transition synthesizer. In the another embodiment, videos in the database are annotated by their transition free video subsequences, shots. This information is provided to avoid the transition synthesizer from accidentally using two video sequences that already contain transition effects. Such a sample would be an outlier in the training set.

[0025] In one embodiment a video database can be approximated by adding only videos to the database for which transitions besides hard cuts and fades are rare. Various shot detection algorithms can perform hard cut and fade detection reliably in order to pre-segment the videos and generate the annotations automatically. The probability that a few complex transition effects would be chosen to produce a sample transition is very rare and can thus be ignored.

[0026] The transition synthesizer is to generate a random video containing the specified number of transition effects of the specified kind. In one embodiment, the following parameters are given before the synthetic transitions can be created:

N = Number of transitions to be generated

$P_{TD}(t)$ = Probability distribution of the duration of the transition effect

R_f, R_b = Amount of forward and backward run before and after the transition. Usually, R_f, and R_b will be set to the same value.

[0027] Figure 2 visualizes the various parameters of the transition generation synthesizer according to one embodiment of the invention as follows:

(1) Read in the list of all videos in the database together with their shot description.

```
(2) For i = 1 to N
    (2.1) Randomly choose the duration d of the transitions according to $P_{TD}(t)$
    (2.2) Determine the minimal required duration for both shots as $(d + R_f)$ and $(d + R_b)$, respectively.
    (2.3) Randomly choose both shots $S1=[t_{s1},t_{e1}]$ and $S2=[t_{s2},t_{e2}]$ subject to their minimal required duration.
    (2.4) Randomly select the start time $t_{start1}$ and $t_{start2}$ of the transition for S1 and S2 subject to $t_{s1}+R_f < t_{start1}< t_{e1} - d$ and $t_{s2} < t_{start2} < t_{e2} - R_b - d$.
    (2.5) Create the video sequence as $S1(t_{start1} - R_f, t_{start1}) + \text{Transition}(S1(t_{start1}, t_{start1}+d), S2(t_{start2}, t_{start2}+d)) + S2(t_{start2}+d, t_{start2}+d+R_b)$.
```

[0028] In one embodiment the transition effect detection system relies on the fixed-scale, fixed position transition detector developed in the training system. More specifically, a fixed location and fixed duration dissolve classifier is developed where dissolves at different locations and of different duration are detected by re-scaling the time series of frame-based feature values and evaluating the classifier at every location in between two hard cuts.

[0029] Figure 3 illustrates a system overview of a transition detection system using a multi-resolution approach according to one embodiment of the invention. First, various frame-based features are derived (figure 3(a)). Each frame-based feature forms a time series, which in turn is re-scaled to a full set of time series at different sampling rates creating a time series pyramid (figure 3(b)). At each scale, a fixed-size sliding window runs over the time series, serving as the input to a fixed-scale and fixed-position transition detector (figure 3(c)). The fixed-scale and fixed position transition detector outputs the probability that the feature sequence in the window belongs to a transition effect. This results in a set of time series of transition effects probabilities at the various scales (figure

3(d)). For scale integration, all probability times series are rescaled to the original time scale (figure 3(e)), and then integrated into a final answer about the probability of a transition at a certain location and its temporal extend (figure 3(f)).

[0030] The computational complexity as well as the performance can be improved by specialized pre- and post-filters. The main purpose of the pre-filter besides reducing the computational load is to restrict the training samples to the positive examples and those negative examples which are more difficult to classify. Such a focused training set usually improves the classification performance.

[0031] Figure 4 illustrates a typical time series of the edge strength feature according to one embodiment of the invention. Edge-based Contrast (EC) captures and amplifies the relation between stronger and weaker edges. In Figure 4, the time series of our dissolve features almost always exhibit a flat graph. Exceptions are sections with camera motion and/or object motion. Thus, the difference between the largest and smallest feature value in a small input window center around the location of interest is used for pre-filtering. If the difference is less than a certain empirical threshold the location will be classified as non-dissolve and is not further evaluated. For multi-dimensional data, the maximum difference between the maximum and minimum in each dimension is used as the criterion. In one embodiment, the input window size is empirically set to 16 frames.

[0032] Figure 5 illustrates the performance of the various features for pre-filtering according to one embodiment of the invention. In general, contrast-based and color-based features respond sometimes differently to typical false alarm situations. Thus, using both kind of features jointly helps to reduce the false alarm rate.

[0033] Figure 5 shows the percentage of falsely discard dissolve location (x-axis) versus the percentage of discard locations (y-axes). Here, the window size was 16 frames and the data has been derived from our large training video set. As can be seen from Figure 5, the YUV histograms outperformed the other features. In this embodiment, a 24 bin YUV image histogram is used (8 bin per channel, each channel separately) to capture the temporal development of the color content.

[0034] Combining YUV histograms with contrast strength (CS) by a simple OR strategy (one of them has to reject the pattern), performs even better, and is chosen as the pre-filter in one embodiment. Generally, the image contrast decreases towards the center of a dissolve and recovers as the dissolve ends. This characteristic pattern can be captured by the time series of the average contrast of each frame. The average contrast strength is measured as the magnitude of the spatial gradient, i.e.,

$$CS_{avg}(t) = \frac{\sum_{x \in X} \sum_{y \in Y} \left\| \left(\frac{\partial}{\partial x} I(x,y,t), \frac{\partial}{\partial y} I(x,y,t) \right) \right\|_2}{|X||Y|}$$

[0035] For simplicity, also the sum of the magnitude of the directional gradients can be used:

$$CS_{avg}(t) = \frac{\sum_{x \in X} \sum_{y \in Y} \left| \frac{\partial}{\partial x} I(x,y,t) \right| + \left| \frac{\partial}{\partial y} I(x,y,t) \right|}{|X||Y|}$$

[0036] However, both of these equations for contrast strength are merely examples and others could be used without departing from the invention.

[0037] In another embodiment, the missed rate of accidentally discarded dissolve locations is set to 2%. Note, since dissolves last many frames, discarding 2% of the dissolve locations must not necessarily result in any loss of a dissolve, especially since in one embodiment the fixed-scale and fixed-position classifier is trained to respond not just to the center of a dissolve, but to the four most centered locations. Regardless, the invention is not limited to discarding 2% and other percentages could be used.

[0038] Given a 16-tap input vector from the time series of feature values, the fixed scale transition detector classifies whether the input vector is likely to be calculated from a certain type of transition lasting about 16 frames (other embodiments may use a varying number of frames without varying from the essence of the invention). There exist many different techniques for developing a classifier. In the following embodiment, a real-valued neural network with hyperbolic tangent activation function is used with the size of the hidden layer as four, which in turn is aggregated into one output neuron. The value of an output neuron can be interpreted as the likelihood that the input pattern has been caused by a dissolve. However, it should be understood that any kind of machine learning technique could be applied here such as support vector machines, Bayesian learning, and decision trees, or Linear Vector Quantizer (LVQ).

[0039] In one embodiment for training and validation, each 10 hours of dissolve videos is synthesized with 1000 dissolves, each lasted 16 frames. The four 16-tap feature vectors around each dissolve's center are used to form the dissolve pattern training/validation set. All other patterns, which do not overlap with a dissolve and are not discarded by the pre-filter, form the non-dissolve training/validation set. Thus, in this embodiment each training and validation set will contain 4000 dissolve examples, and about 20000 non-dissolve examples.

[0040] Figure 6 is a block diagram further illustrating the creation of the training and validation set of block 100 according to one embodiment of the invention. In block 110, the transition effect type and its desired parameter distribution are set. If a training set is to be created then control passes to block 120 from block 110. If a validation set is to be created then control passes to block 130.

[0041] In block 120, the system creates a long training video sequence with a given number of transitions and control passes to block 140. In block 140, the feature values are derived, the training samples are created and added to the training set. Control is then passed to block 160. In block 160, the training set is outputted.

[0042] In block 130, the system creates a long validation video sequence with a given number of transitions and control passes to block 150. In block 150, the feature values are derived, the training samples are created and added to the training set. Control is then passed to block 170. In block 170, the training set is outputted.

[0043] Initially 1000 dissolve patterns and 1000 non-dissolve patterns are selected randomly for training. Only the non-dissolve pattern set is allowed to grow by means of the so-called 'bootstrap' method, although other embodiment may use techniques other than the bootstrap method. This method starts with training a neural network on the initial pattern set. Then, the trained network is evaluated using the full training set. Some of the falsely classified non-dissolve patterns of the full training set are randomly added to the initial pattern set and a new, hopefully enhanced neural network is trained with this extended pattern set. The resulting network is evaluated with the training set again and additional falsely classified non-dissolve patterns are added to the set. This cycle of training and adding new patterns is repeated until the number of falsely classified patterns in the validation set does not decrease anymore or nine cycles has been evaluated. Usually between 1500 and 2000 non-dissolve pattern may be added to the actual training set. The network with the best performance on the validation set is then selected for classification. Figure 7 further illustrates this process. Note that in other embodiments of the system, falsely classified dissolve and non-dissolve patterns are added to the pattern set, not just falsely classified non-dissolves patterns.

[0044] Figure 7 is a block diagram further illustrating the detector training of block 200 according to one embodiment of the invention. In block 210, X_1 positive and X_2 negative training examples are taking as current training sets, then control passes to block 220. In block 220, a run count is set to 1, then control passes to block 230. In block 230, a new neural network is trained with the current training set, then control passes to block 240. In step 240, the trained neural network is used to classify all training patterns. A small number of falsely classified patterns are randomly selected and added to the current training set. Control then passes to block 245. In block 245, if the maximum run count is not reached then control passes back to block 230. However, if the maximum run count is reached then control passes to block 250. In block 250, all classifiers are validated and the neural network with the best performance on the validation set is chosen as the fixed-scale fixed position detector in detection system. In block 260, the best neural network is outputted.

[0045] A problem that may be encountered by any dissolve detection method is that there exist many other events that may show the same pattern in the feature's time series. Therefore, in order to reduce the false hits in one embodiment, a restriction is made to detect type one dissolves during post-filtering and, thus check for every detected dissolve whether its boundary frames qualify for a hard cut after its removal from the video sequence. If it does not qualify, then the detected dissolve is discarded.

[0046] In addition, in one embodiment it is assumed that the dominant camera motion operation from the video are caused by pans and zooms as determined by the number of false alarms. Thus, all detected dissolves which temporally overlap by more than a specific percentage with a strong dominant camera motion are also discarded during post-filtering. In one embodiment, all detected dissolves which temporally overlap by 70% are discarded.

[0047] These two post-filtering criteria help to reduce the false alarm rate and are applied on each scale. In the present embodiment, the output of the post-filtering stage is a list of dissolves with the following parameters: <scale><from><to><prob(dissolve)>.

[0048] It is important to note that the fixed-scale and fixed position transition detector may be very selective. That is, it might only respond to a dissolve at one scale. Therefore, in another embodiment a winner-takes-all strategy may be implement. Here, if two detected dissolve sequences overlap, then the one with the highest probability value wins (i.e., the other is discarded). The competition starts at the smallest scale (short dissolves) competing with the second smallest scale and goes up incrementally to the largest (long dissolves).

[0049] Wherein embodiments have described in which the transition type “dissolve” is used to demonstrate the new detection system, alternative embodiments could be implemented to demonstrate the invention with other transition types or special effects in videos.

[0050] Also wherein embodiments have described in which a neural network classifier is used to demonstrate the new detection system, alternative embodiments could be implemented to demonstrate that a classifier based on other machine learning algorithms such as support vector machines, Bayesian learning, and decision trees could be used instead.

[0051] While the invention has been described in terms of several embodiments, those skilled in the art will recognize that the invention is not limited to the embodiments described.

[0052] The method and apparatus of the invention can be practiced with modification and alteration within the spirit and scope of the appended claims. The description is thus to be regarded as illustrative instead of limiting on the invention.